Identifying Salient Entities in Web Pages

Michael Gamon
Microsoft Research
One Microsoft Way
Redmond, WA, USA
mgamon@microsoft.com

Tae Yano
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
taey@cs.cmu.edu

Xinying Song
Microsoft Research
One Microsoft Way
Redmond, WA, USA
xinson@microsoft.com

Johnson Apacible
Microsoft Research
One Microsoft Way
Redmond, WA, USA
johnsona@microsoft.com

Patrick Pantel
Microsoft Research
One Microsoft Way
Redmond, WA, USA
ppantel@microsoft.com

ABSTRACT

We propose a system that determines the *salience* of entities within web documents. Many recent advances in commercial search engines leverage the identification of entities in web pages. However, for many pages, only a small subset of entities are central to the document, which can lead to degraded relevance for entity triggered experiences. We address this problem by devising a system that scores each entity on a web page according to its centrality to the page content. We propose salience classification functions that incorporate various cues from document content, web search logs, and a large web graph. To cost-effectively train the models, we introduce a *soft labeling* methodology that generates a set of annotations based on user behaviors observed in web search logs. We evaluate several variations of our model via a large-scale empirical study conducted over a test set, which we release publicly to the research community. We demonstrate that our methods significantly outperform competitive baselines and the previous state of the art, while keeping the human annotation cost to a minimum.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Document aboutness, entity salience, content analysis.

1. INTRODUCTION

The concept of *Salience* or *Aboutness* has been investigated in many fields of research, from linguistics to semiotics, and from sociology to psychology. While dictionary definitions look deceptively simple ("most noticeable or important" (OED), "state or condition of being prominent" (Wikipedia)), the notion of salience is very hard to pin down in practice. A number of observations around salience might be uncontroversial, however: (1) Salience and relevance/importance are not the same. An entity or notion A in a text can be highly salient, yet unimportant to the reader. (2) Salience is a function of the structure of a text, and indirectly a function of the intention of the author, as opposed to a function of the reader's intent or needs.

Salience also has very practical implications on the Web: People, entities, and content are increasingy linked in a "Web of Things" paradigm [5]. However, in our samples we found that fewer than 5% of entities on a page are salient to the web page, making it very important to be able to distinguish them from the remaining non-salient entities.

We propose scalable weakly-supervised models for learning to score entities according to their salience to a document. We leverage web search logs to automatically acquire *soft labels* as a supervision signal for our training data. We train our models on a large number of web pages, leveraging features from document content, page classifiers, and a web graph. Finally we show empirical evidence, on data representing the `HEAD` and `TAIL` distributions of the web, that our methods significantly outperform the previous state of the art on various ranking and classification metrics. As this is the first dataset created of its kind, we release it publicly to the research community.

The major contributions of this paper are:

- We devise a notion of entity salience and frame the problem of understanding the aboutness of a document as determining the most salient entities in the document;
- We model the task of entity salience detection as a weakly supervised machine learned model, generating labeled training data via usage behaviors found in web search logs;
- We present empirical evidence that our system significantly outperforms previously established baselines;
- We publicly release test sets consisting of URLs and their entity mentions randomly drawn from head and tail distributions in a commercial web search engine along with gold standard salience judgments.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505602.

2. RELATED WORK

Understanding the meaning or aboutness of a document has received attention from both a theoretical [23, 12, 2] and practical perspective. In the latter approaches, driven by application-specific demands, computational models have decomposed aboutness and focused on detecting aspects of aboutness such as key terms [30, 13, 21], latent semantic spaces/topics [18, 1], and summaries [25, 17, 10].

Most related to our work is the current state of the art described in Paranjpe [21] where the focus is on the detection of key terms in web pages. There are three key differences between our proposed method and theirs. First, our notion of document aboutness is entity-centric, i.e., we consider the identification of salient entities, as opposed to salient terms of any kind. Second, our soft labeling method is different from theirs and we demonstrate that it outperforms it and is more robust to effects of popularity and presentation order of URLs in the search results page (SERP). Finally, our feature set is a significant extension of the set of features Paranjpe utilizes.

The keyword extraction task can be seen as related to entity salience, where keywords and key phrases are a superset of salient entities in a document. Keyword extraction is often addressed in the context of various document understanding tasks, most often in extraction based summarization or abstract generation [11, 20, 10], and more recently in (online) contextual advertisement or keyword appraisal [30]. Linguistic cues, including syntactic, semantic and discourse information for keyword extraction are investigated in [11], [20], [12], [8] and [2], for example.

Term frequency statistics and term weighting schemes are also commonly used to score the specificity/importance of a term in information retrieval [19, Chapter 7]. The same idea motivates the vector space model and its application in ad hoc document retrieval, indexing and key phrase generation [27, 26].

We use supervised machine learning to build our models of entity salience, a method that has been used widely for various tasks in web document processing. Machine learning offers a principled way to calibrate signals from heterogeneous sources, which is crucial when incorporating diverse (e.g. document content, term-weighting, web graph) insights into one system. A wide variety of tactics are employed in the literature to overcome the bottleneck of acquiring supervision data, for a theoretical perspective on these approaches see [31]. One well studied approach to obtain relevance-related supervision for web document training data is the use of web search logs: the click behavior that is recorded in these logs can serve as implicit user feedback and hence indicate relevance of a document to a user. This signal has been exploited for relevance annotation in document retrieval systems [14, 15, 9, 24], for other web document tasks beyond retrieval, i.e., [29], [16], [22], and [13]. In our system, we exploit web search logs by designing a soft labeling function for entity salience that is based on user behavior information.

3. ENTITIES, SALIENCE AND WEB PAGES

3.1 Entity Salience

What constitutes an entity has been cause of many philosophical debates. For our purposes, we consider something an entity if it is of a type that has or reasonably could have a Wikipedia page associated with it. This would include people, places, companies, as well as events, concepts, and famous dates.

We consider the following working assumptions in building our entity salience ranking system:

- **Local scoping:** The salience of an entity can be solely determined by how the entity is presented within the document. In other words, entity salience can be effectively computed from the local context, or what is available in the document itself.
- **Invariable perception:** Entity salience can be assessed independently from the intentions or interests of a user/reader, and independently from the prior importance of the entity as it exists outside of the document.

Entity salience is distinct from two other aspects of aboutness: *entity importance* and *entity relevance*. The importance of an entity refers to its influence or substantiveness outside of the scope of the document. For example, although *Barack Obama* is a very important entity, he can be peripheral to some news stories. On the other hand, the relevance of an entity is inherently subjective to the reader's perspective and intent.

Although local scoping suggests that the evidence for entity salience can be derived most effectively from the document content, it is important to note that extra-document information such as incoming anchor links and user click-through data provide important signal, and will be leveraged by our models. Also, by assuming the source of salience to be local to a document, we limit the search space to those entities in the document.

3.2 Salient Entities and Web Pages

We conducted a small manual inspection of web pages in order to get a first perspective at the difficulty and scope of our problem. We sampled 50 documents, randomly chosen from a traffic-weighted sample of documents from a commercial web search index. We examined the content of the pages in a web browser and made a list of all entities and their salience.

On average, fewer than 5% of the entities in each document were deemed salient. We observed certain cues when identifying salience. Unsurprisingly, salient entities tend to be mentioned in the title, headings, and/or first paragraph, and are frequently mentioned.

By our local scoping assumption, any salient entity is contained in its document. Hence, a system that is capable of identifying each entity in a document would serve as a candidate generator for a salience ranking system. We ran a proprietary state-of-the-art NER system, trained using the perceptron algorithm [4], on the content of the web documents. We then compared for each page the set of automatically identified entities to the human annotation.

We found that in 91% of the documents, at least one of the salient entities is in the candidate entity set identified by our NER tagger. For over 90% of these pages, all the human annotated salient entities are captured by our NER engine. Therefore it is reasonable to use the NER system as a candidate generator.

We next examined whether simple cues for entity salience are so straightforward that a heuristic would suffice to identify them. We observed many cases where cues were not reliable or conflicted with each other, making heuristic de-

sign a difficult proposition. For example, the presence of an entity in a title string is often a good indicator for salience. However, being included in the title (or in the first paragraph) is neither a necessary nor a sufficient condition for salience. Based on these observations, we believe that a machine learned model that can combine evidence from a multitude of signals is a better approach than developing simple heuristics.

4. MODEL

4.1 Task Definition

Let $\mathbf{D}$ and $\mathbf{E}$ be the sets of all documents and entities on the web, respectively. Let $\mathbf{E_d} \subset \mathbf{E}$ be the set of entities mentioned in $d \in \mathbf{D}$. We formally define the aboutness task as learning the function:

$$\sigma : \mathbf{D} \times \mathbf{E} \rightarrow \mathbb{R} \quad (1)$$

where $\sigma(d, e)$ reflects the salience of e in d[1].

We denote the ranking of $\mathbf{E_d}$ according to σ as:

$$\mathbf{R_d^S} = (e_1, ..., e_{|E_d|} \mid e_i \in E_d, \sigma(d, e_i) \geq \sigma(d, e_{i+1}))$$

where pairs of entities with tied scores are ordered randomly. We define the ranking function

$$R_\sigma : \mathbf{D} \times \mathbf{E} \rightarrow \mathbb{N} \quad (2)$$

such that $R_\sigma(d, e)$ equals the rank of e in $\mathbf{R_d^S}$[2].

4.2 Soft Labeling

Instead of manually labeled data we rely on a *soft labeling* approach that uses behavioral signals from web users as a proxy for salience annotation. Individual clicks in a web search log from a commercial search engine indicate a user's interest in a URL based on their entity query, i.e., they indicate the relevance of the entity in the URL to the user. In aggregate, the combined interests for an entity/URL pair will correlate with the entity being salient, since users are less likely to search for an entity and then examine a page that is not about that entity. This "soft label" is available for pages that receive enough traffic to derive reliable user click statistics, but the learned model uses features that are independent of user behavior, hence it can generalize to the tail of the distribution.

A simple click measure is Clickthrough Rate (CTR), i.e. the rate at which users click on a URL given a query. Paranjpe [21] points out that CTR is very much biased towards the top-ranked result on the SERP which tends to receive the bulk of user clicks. Instead, they propose to use Click Attractivity (CA) as a search log based metric that correlates with salience. CA for a term t and document d is defined as:

$$CA(t, d) = \frac{clicks(t, d)}{clicks(t, d) \ + \ skips(t, d)} \quad (3)$$

where $clicks(t, d)$ is the number of times users clicked on d for a query containing t, and $skips(t, d)$ is the number of times users clicked on another document d' that is ranked at a lower position than d, where d is in the top-5 results. Both *clicks* and *skips* are aggregated over all queries that contain t and lead to at least 32 instances where document d is displayed in the SERP. In its original setting, CA was used for any term t in a document. For this paper, only terms that are entities are considered, i.e., $t = e$.

CA has the following problems, though: First, recency can trump salience. Assume that entity e is involved in some recent gossip news. The user will be most interested in the latest gossip about e (which provides a good signal for salience) but will hardly ever click on the IMDB or Wikipedia page for e, although on these pages e is very salient. Second, popularity can trump salience. Within a set of URLs that are equally about e, some of the sites might be more popular than others (e.g., a celebrity home page will be more popular than a page about her maintained by a fan.) This will distort the CA score. Finally, CA is subject to position bias similarly to CTR. If the user is generally more likely to click on a URL in the top position, this also means that she is less likely to skip that top position and hence that CA is also influenced by position bias.

We propose a different soft labeling function that aggregates over only the queries that lead to clicks on a URL without taking the number of views (CTR) or the number of skips (CA) into account. We define Entity Query Ratio (EQR) for entity e and document d by looking at all queries that lead to a click on d. Within that set of queries, we calculate the ratio of the number of clicks from queries containing e to the number of clicks from all queries. We define *containing* here as an exact match between an entity string and a query. In our experiments, this strict matching definition performed better than a substring-based definition.

$$EQR(e, d) = \frac{clicks(e, d)}{\sum_{q \in Q} clicks(q, d)} \quad (4)$$

where Q is the set of all queries and $clicks(e, d)$ is redefined as the number of times users clicked on d for a query matching e.

4.3 Features and Learning Algorithm

We represent each entity/document pair $\langle e, d \rangle$ as a vector of features. At the highest level, there are three distinct classes of features: (1) those that are computed from properties of e and the whole document collection D, labeled $\mathrm{F}_{e,\mathbf{D}}$; (2) those that are solely computed from properties of d, labeled F_d; and (3) those that are computed from properties of e in d, labeled $\mathrm{F}_{e,d}$. Document features, F_d, further subdivide into categorical features representing the page classification of d, features of the document URL, and length features. Entity/document features, $\mathrm{F}_{e,d}$, are subcategorized into structural features that relate e to the structure of d, web graph features that indicate the frequency of e in inlinks and outlinks, position features that capture the location of e in d, and finally features that capture the frequency of e in 17 different page segments that are automatically identified based on visual properties (see [3] for more details and [28] for other work that used visual blocks as input). Novel features considered in this paper include: Page classification and segmentation features, more detailed position features, and corpus features based on an offline corpus of documents in the top domain of d.

We use regression and ranking learning to model CA and EQR. We employ boosted decision trees [6] as our learning algorithm. The hyperparameters are the number of iterations, learning rate, minimum instances in leaf nodes, and

[1] We fix $\sigma(d, e) = 0$ for all $e \notin \mathbf{E_d}$.

[2] $R_\sigma(d, e)$ is not defined for $e \notin \mathbf{E_d}$.

the number of leaves. The parameter tuning procedure is described in Section 6.1.

5. EVALUATION METHODOLOGY

5.1 Test Set Construction

Let ρ be a graded relevance scoring function for a document d and entity e:

$$\rho : \mathbf{D} \times \mathbf{E} \rightarrow \{\mathbf{MS}, \mathbf{LS}, \mathbf{NS}\} \quad (5)$$

where for $\rho(d, e)$:

- **Most Salient (MS)** indicates that d is mostly about e, or e plays a prominent role in the content of d;
- **Less Salient (LS)** indicates that e plays an important role in some parts of the content of d; and
- **Not Salient (NS)** indicates that d is not about e.

We define a test set $\mathbf{T} = \{\rho, \mathbf{\Delta}\}$ where $\mathbf{\Delta} = \{(d, \mathbf{E_d}) : d \in \mathbf{D}, \mathbf{E_d} \subset \mathbf{E}\}$ is a collection of pairs of web pages and entities for which we have a gold standard ρ.

We start by constructing a universe of web pages by mining all the shared URLs on the full firehose of Twitter.com during May 2012 (to ensure that we focus on URLs that are actively shared and discussed). This set was narrowed down by eliminating: (1) any URL that redirected to a query on a search engine, (2) YouTube.com links (since salient entities here are often trivial to identify), and (3) URLs that received fewer than three clicks within six months. The final set consists of over half a million URLs, for which we have access to a full crawl of the content.

From this set of web pages, we produce 2414 manually annotated test cases for our experiments, spanning two test sets outlined below. Each test set consists of randomly sampled web pages such that each page contains fifty or fewer entities to facilitate manual annotation. The first set, labeled `HEAD`, consists of a traffic-weighted random sample of web pages from our universe of URLs, where the traffic weights are estimated using the number of clicks each URL received during a six month period. This set represents the head distribution of our URLs. The second test set, `TAIL`, consists of a uniform random sample of web pages from our universe of URLs. This set represents the long tail of the web.

For each web page in our test sets, we built the set of entity mentions by running the Named Entity Recognizer, described in Section 3.2, on the content of each page. There are 1228 candidate entities in the `HEAD` set and 1186 in the `TAIL` set. To complete `HEAD` and `TAIL`, we construct gold standard relevance assessments, ρ, for each entity-document pair. We used a crowdsourcing tool to collect relevance judgments (**MS**, **LS**, or **NS**) from non-expert paid judges. For each entity-document pair, we requested five judgments. We removed all judgments from *bad* judges, which were identified as those whose mean judgment score was further than two standard deviations from the mean of all judges. This resulted in the removal of four judges for `HEAD` and seven for `TAIL`. The task had fair agreement for both test sets, with a Fleiss' κ score of 0.29 on `HEAD` and 0.25 on `TAIL`. Three expert judges then adjudicated the majority vote for each entity-document pair.

The `HEAD` and `TAIL` test sets along with their gold standard annotations are available at http://research.microsoft.com/research/downloads/details/5a2ddfde-83f7-4962-9ad7-d80cd5098f38/details.aspx.

5.2 Performance Metrics

To assess the quality of a salience function σ on a test set $\mathbf{T}$, we compute the aggregate performance against the salience judgements given by the human judges. We consider two types of applications. First, rank-sensitive applications, such as those deriving relevance features for a search ranking function, require the top-K most salient entities. For these, classic IR metrics such as **nDCG** (normalized discounted cumulative gain) and **MAP** (mean average precision) are applicable [19]. Second, in class-sensitive applications, such as highlighting the salient entities on a document, we require all the salient entities on the page. For this class of applications, **Precision**, **Recall**, and **F1** metrics are applicable.

Below we define **nDCG** and **MAP** with respect to a system σ, its corresponding ranking function R_σ (Eq. 2), and test set $\mathbf{T}$.

$$\mathbf{nDCG_T}(\sigma) = \frac{1}{\mid \mathbf{T} \mid} \times \sum_{(d, \mathbf{R_d^\sigma}) \in \mathbf{T}} \frac{\sum_{r=1}^{|\mathbf{R_d^\sigma}|} \frac{2^{\phi_{tri}(d, e_r)} - 1}{\log_2(r+1)}}{IDCG(d, \mathbf{E_d})}$$

where $\phi_{tri}(d, e_r)$ maps the relevance score of e_r in d to a real-valued score (**MS** $\rightarrow$ 1.0, **LS** $\rightarrow$ 0.5, **NS** $\rightarrow$ 0) and $IDCG(d, \mathbf{E_d})$ is the ideal DCG if $\mathbf{E_d}$ was perfectly ranked.

$$\mathbf{MAP_T}(\sigma) = \frac{1}{\mid \mathbf{T} \mid} \times \sum_{(d, \mathbf{R_d^\sigma}) \in \mathbf{T}} \frac{\sum_{r=1}^{|\mathbf{R_d^\sigma}|} \phi_{bin}(d, e_r) Prec(\mathbf{R_d^\sigma}[\mathbf{1}, \mathbf{r}], d)}{\sum_{r=1}^{|\mathbf{R_d^\sigma}|} \phi_{bin}(d, e_r)}$$

where $\mathbf{R_d^\sigma}[\mathbf{1}, \mathbf{r}] = \{e_1, ..., e_r | e_i \in R_d^\sigma\}$, $\phi_{bin}(d, e_r)$ indicates if the entity at rank r is salient or not in d, and:

$$\mathbf{Prec}(\mathbf{R}, d) = \frac{\sum_{r=1}^{|\mathbf{R}|} \phi_{bin}(d, e_r)}{\mid \mathbf{R} \mid}$$

Recall and **F1** follow trivially.

6. EXPERIMENTAL RESULTS

6.1 Experimental Setup

We first ran our NER system on the content in our Web Page Data, discarding those pages in our `HEAD` and `TAIL` test sets, and associated with these pages all queries from the US English market of Bing.com that led to a click on the pages during a six month period. We computed the CA and EQR scores for each entity. Many entity-URL pairs receive a zero score because no query mentioning the entity leads to any click on the URL. Although such an entity-URL pair could in fact be salient (even with six months of web search log data, there is sparsity in the tail), in most cases the pair is non-salient. In our experiments, we tried configurations that included all zero-scoring entity-URL pairs, none of them, and balancing the number of zero-scoring pairs to be equal to the number of non-zero-scoring pairs via random sampling. The balanced configurations consistently and by a large margin outperformed the others, and hereon we consider only balanced configurations. For the EQR soft label, our final training set contains 66,055 entity-URL pairs; for the CA soft label the number of entity-URL pairs in the training set is 48,759[3].

[3]This discrepancy in number of training cases is due to the fact that we only compute the CA label for documents in the top 5 displayed search results, to keep the CA signal sufficiently reliable.

	HEAD					TAIL				
	nDCG@1	**nDCG@5**	**MAP@1**	**MAP@5**	**F1**	**nDCG@1**	**nDCG@5**	**MAP@1**	**MAP@5**	**F1**
$\mathbf{CA_{base}}$	0.49	0.54	0.28	0.33	0.55	0.43	0.46	0.27	0.32	0.42
$\mathbf{EQR_{base}}$	0.51	0.54	0.20	0.28	0.55	0.43	0.46	0.12	0.27	0.42
CA_TFIDF	0.66	0.73	0.38	0.46	0.63	0.54	0.57	0.29	0.38	0.48
CA_PJP	0.70	0.80	0.42	0.51	0.66	0.60	0.65	$0.35^{\dagger}$	0.47	0.55
CA_ALL_RANK	$0.80^{\dagger}$	$\mathbf{0.85}^{\ddagger}$	$0.52^{\dagger}$	0.57	0.70	$0.73^{\ddagger}$	$0.72^{\dagger}$	$0.48^{\dagger}$	$0.54^{\dagger}$	$0.59^{\dagger}$
CA_ALL	$0.80^{\dagger}$	$\mathbf{0.85}^{\dagger}$	$0.52^{\dagger}$	$0.57^{\dagger}$	0.70	$0.65^{\dagger}$	$0.76^{\ddagger}$	$0.40^{\dagger}$	$0.54^{\dagger}$	$0.61^{\dagger}$
EQR_TFIDF	0.60	0.71	0.32	0.43	0.59	0.56	0.58	0.31	0.40	0.49
EQR_PJP	$\mathbf{0.82}^{\ddagger}$	0.81	$\mathbf{0.54}^{\ddagger}$	$0.56^{\dagger}$	0.69	$0.65^{\dagger}$	$0.66^{\dagger}$	0.40	0.46	0.56
EQR_ALL_RANK	$0.80^{\dagger}$	$0.84^{\dagger}$	$0.52^{\dagger}$	$\mathbf{0.58}^{\dagger}$	$\mathbf{0.75}^{\dagger}$	$\mathbf{0.77}^{\ddagger}$	$0.74^{\dagger}$	$\mathbf{0.52}^{\ddagger}$	$0.56^{\dagger}$	$0.59^{\dagger}$
EQR_ALL	$\mathbf{0.82}^{\ddagger}$	$\mathbf{0.85}^{\dagger}$	$\mathbf{0.54}^{\ddagger}$	$\mathbf{0.58}^{\dagger}$	$\mathbf{0.75}^{\dagger}$	$0.73^{\ddagger}$	$\mathbf{0.77}^{\ddagger}$	$0.48^{\ddagger}$	$\mathbf{0.58}^{\ddagger}$	$\mathbf{0.64}^{\ddagger}$

Table 1: Model analysis on HEAD and TAIL against rank-sensitive metrics (nDCG and MAP) and classification-sensitive metric F1. † indicates statistical significance over the soft labeling baselines and the tf.idf feature configuration; ‡ further indicates statistical significance over CA_PJP (significance assessed using Student's t-Test with p-value = 0.1). Bold indicates the highest achieved score on each metric.

For each soft-labeled entity-URL pair, we computed the features described in Section 4.3. We used the Bing search engine to compute features that require web graph data or page classification. To set the hyperparameters of our regression and ranking models from Section 4.3, we perform a sweep of 144 combinations of parameter settings on a three-fold cross validation, for each system configuration.

Each system that we train and evaluate consists of three choices: soft labeling method (CA vs. EQR), feature set, and model type (regression and ranking).

We consider the following five baselines against which to test our systems:

- $\mathbf{CA_{base}}$ and $\mathbf{EQR_{base}}$: The systems that use the CA and EQR soft labels as their prediction (without a learned model);
- **CA_TFIDF** and **EQR_TFIDF**: Regression models using only the `tf`, `df`, `tf.idf` features.
- **CA_PJP**: Current state-of-the-art model [21].

We report our results on the following system configurations:

- **EQR_PJP**: Regression model with the feature set from [21] with our soft labeling function.
- **CA_ALL** and **EQR_ALL**: Regression models with all features.
- **CA_ALL_RANK** and **EQR_ALL_RANK**: Ranking models with all features.

6.2 System Comparison

Table 1 lists the performance of our baseline and system configurations on both the HEAD and TAIL datasets. We report nDCG and MAP scores (at 1 and 5) and F1. **EQR_ALL** and **EQR_ALL_RANK**, our best configurations, significantly outperform the soft labeling baselines, on both HEAD and TAIL, by 37% and 51% on F1, respectively. On TAIL, we improve on the previous state of the art, **CA_PJP**, significantly on all metrics, by 16% on F1. On HEAD, we show significant improvement over **CA_PJP** in the first position on both nDCG and MAP.

In general, the HEAD is "easier" than the TAIL: Absolute metrics are higher, and the choice of feature sets and soft labeling function matters less. This is not surprising for two reasons: (1) the soft label signal is reliable only in the head but it is extremely sparse in the tail; and (2) the head is represented dominantly in the training data. As [21] points out, the strategy behind learning a salience model from a soft label is to learn from the cases where we have a good supervision signal and to generalize to the cases in the tail. Given this argument, our expectation was to see gains mostly in the tail for our proposed soft labeling function and feature set **EQR_ALL**. The positive gains on the HEAD were unexpected. The choice of the soft labeling function is important: On TAIL, EQR outperforms CA overall as a training signal. On HEAD, the soft labeling technique matters less; using our full feature set, both techniques yield similar performance except on F1 where EQR outweighs CA. Using our rank models, we observe par performance against the regression models on HEAD. On TAIL, the rank models outperform regression in the first position on nDCG and MAP.

Examining the precision/recall characteristics of the systems, we found that the TFIDF features underperform compared to the other feature sets in all settings. The PJP feature set improves precision/recall in all cases against TFIDF, but in a more pronounced fashion when used with the EQR soft labels. The best precision/recall curves are obtained from **EQR_ALL**. The best system produces precision/recall gains especially in the region where precision is greater than 0.7. At recall ~ 0.6 the precision gain on HEAD is nearly 7.5 points, on tail it is nearly 10 points at ~ 0.5 recall.

6.3 Contribution of Feature Families

Examination of the feature weights in **EQR_ALL** reveals that the strongest salience cues are the position and the frequency of the entity in the document and anchor text. In the model, 174 features receive non-zero weights. The top five features are: the frequency of e in the anchor text, document and title, and the `df` of e and offset of e in the document. The next series of 37 features in order of feature weight is a mix of page classification, position, URL, structural and page segmentation features with no discernible prominence of any of these families. The binary features representing top level domains and page categories occur in the lower weight area of the feature list, with the exception of the feature indicating that the top level domain is Wikipedia - this feature ranks 11^{th} which is not surprising given the frequency and highly specific structure of this domain. We also performed feature ablation on **EQR_ALL** to see how well a system that does not have access to information that requires either a sizeable web crawl or components that are typically part of a commercial search engine would do. On HEAD,

the difference is minimal and not statistically significant, except for F1 where **EQR_ALL** outperforms **EQR_DOC**. In `TAIL`, however, **EQR_ALL** achieves better results, with significant gains in nDCG@1 and nDCG@5.

7. CONCLUSION

This paper formalizes and addresses the task of scoring entity-URL pairs according to the salience of the entity in the document. We propose a system that is cost-effective to build and improves upon the state of the art. We propose weakly-supervised learned models combined with a novel method for automatically labeling large quantities of training data by leveraging usage behaviors found in web search logs. This, along with an extensive feature set leads to significant improvements over the current state of the art on both head and tail distributions of the web. As no public data exists to date to evaluate this task, we design and release to the research community a gold standard data set with salience annotations, representing the head and tail distributions of pages on the web.

For further details, we refer the reader to [7].

8. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] P. D. Bruza, D. W. Song, and K. F. Wong. Aboutness from a commonsense perspective. *Journal of the American Society for Information Science*, 51:1090–1105, 2000.

[3] D. Cai, S. Yu, J. Wen, and W. Ma. Extracting content structure for web pages based on visual representation. *Web Technologies and Applications*, pages 406–417, 2003.

[4] M. Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*, 2002.

[5] N. N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu. A web of concepts. In *Proceedings of PODS*, 2009.

[6] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 1999.

[7] M. Gamon, T. Yano, X. Song, J. Apacible, and P. Pantel. Understanding Document Aboutness - Step One: Identifying Salient Entities. Technical Report MSR-TR-2013-73, Microsoft Research, 2013.

[8] B. Hjørland. Towards a theory of aboutness, subject, topicality, theme, domain, field, content... and relevance. *Journal of the American Society for Information Science and Technology*, 52(9):774–778, 2001.

[9] S. Holland, M. Ester, and W. Kießling. Preference mining: A novel approach on mining user preferences for personalized applications. *Knowledge Discovery in Databases: PKDD 2003*, pages 204–216, 2003.

[10] E. Hovy and C. Y. Lin. Automated text summarization and the summarist system. In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*, 1998.

[11] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of EMNLP*, 2003.

[12] W. Hutchins. On the problem of 'aboutness' in document analysis. *Journal of Informatics*, 1(1):17–35, 1977.

[13] U. Irmak, V. V. Brzeski, and R. Kraft. Contextual ranking of keywords using click data. In *Proceedings of ICDE*, 2009.

[14] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of SIGKDD*, 2002.

[15] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR*, 2005.

[16] M. Komachi and H. Suzuki. Minimally supervised learning of semantic knowledge from query logs. In *Proceedings of IJCNLP*, 2008.

[17] J. Kupiec, J. O. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of SIGIR*, 1995.

[18] T. Landauer and S. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[19] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[20] D. Marcu. From discourse structures to text summaries. In *Proceedings of ACL*, 1997.

[21] D. Paranjpe. Learning document aboutness from implicit user feedback and document structure. In *Proceedings of CIKM*, 2009.

[22] M. Paşca and B. V. Durme. What you seek is what you get: Extraction of class attributes from query logs. In *Proceedings of IJCAI*, 2007.

[23] H. Putnam. Formalization of the concept 'About'. *Philosophy of Science*, 25(2):125–130, 1958.

[24] F. Radlinski and T. Joachims. Query Chains: Learning to rank from implicit feedback. In *Proceedings of SIGKDD*, 2005.

[25] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of SIGIR*, 1993.

[26] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[27] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[28] R. Song, H. Liu, J. Wen, and W. Ma. Learning block importance models for web pages. In *Proceedings of WWW*, 2004.

[29] G. Xu, S. Yang, and H. Li. Named entity mining from click-through data using weakly supervised latent dirichlet allocation. In *Proceedings of SIGKDD*, 2009.

[30] W. Yih, J. Goodman, and V. Carvalho. Finding advertising keywords on web pages. In *Proceedings of WWW*, 2006.

[31] X. Zhu. Semi-Supervised Learning Literature Survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.